An Empirical Approach for Semi-Supervised Sentiment Analysis and Opinion Mining

Dhanashree Raut[1], Seema Ladhe[2]
[1] *PG Scholar, Computer Engineering, MGMCET, Kamothe, Navi Mumbai (India)*
[2] *Professor, Computer Engineering, MGMCET, Kamothe, Navi Mumbai (India)*

Abstract -Online shopping has become popular as it is convenient, reliable and cost effective. The major challenge in front of the online customers is purchasing decision based on the pictures or description provided online. Reviews of the product often makes it easy for the customers to make decisions for purchasing the products as reviews are a great source to compare products and features. The existing schemes include NLP and statistical methods for opinion mining, document level classification and multi domain sentiment analysis. These schemes are used for formal reviews and to remove the grammatical errors. We have proposed Feature Extraction and Refinement for Opinion Mining (FEROM), a scheme based on the opinion miner system that removes the similar meaning words for informal reviews .FEROM reviews the opinions expressed by the customer and identifies their expressions and opinion orientations for each recognized product entity. Our scheme will positively help customers for taking effective purchasing decision since it results in segregation of positive and negative reviews.

Keywords: Opinion Mining, FEROM, NLP, Feature Extraction, Refinement.

I. INTRODUCTION

Recently, a number of online customers have dramatically increased due to rapid growth of e-commerce. Now-a-days, the merchants and product manufacturers allow customers to post their review or express opinions on products and services on sites such as amazon.com, flipkart.com, eBay. in for enhancement of customer satisfaction.

These online reviews thereafter ,become a collective source of information for both the customers and the product manufacturers. The customers utilize this piece of information to make their purchasing decisions more effective and from the perspective of the product manufacturers preferences of customer is highly valuable for the development and manufacturing of product and also, for the customer relationship management. As far as semi-supervised learning is considered, it suits well for the domains of opinion mining and sentiment analysis because a labeling example is a difficult task. The labeled examples give correct classification, but sometimes unlabelled example give higher classification accuracy than labeled ones as they fill the gaps that exist between the labeled ones.

II. SYSTEM MODEL

Opinions are central to almost all human activities and are key influencers of our behaviours. The choices we make and our perceptions of reality and beliefs depend to a considerable degree, on how others see and assess the world. For this reason, when we need to make a decision we often peruse the opinions of others. This is merely not true for individuals but also for organizations

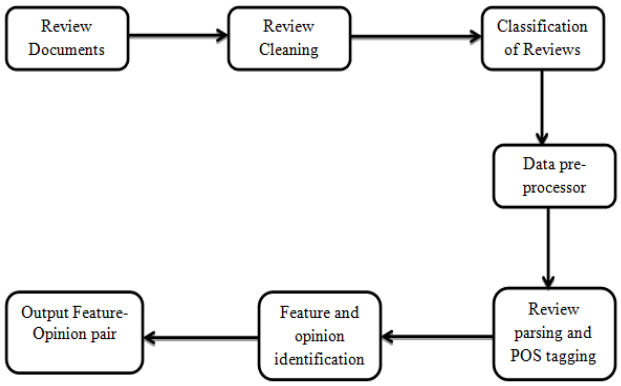

Fig 2.1.Opinion Mining System

Opinion related concept such as sentiments, emotions and evaluations are the subjects of sentiment analysis and opinion mining[1].The speedy growth of this field coincide with those of social media on the web, e.g. ,blogs, reviews[2],forums and social networks[10],because for the first time a enormous volume of opinionated data has been recorded in digital forms. In recent years, sentiment analysis has grown one of the most active research areas in NLP and is also widely studied in data mining, text mining and web mining. Sentiment analysis has also found applications in almost every business and social domain.

A. The Opinion Mining System

The opinion mining system in Fig 2.1 gathers the review from online sites and removes unwanted sentences. It then classifies the reviews and the data in the sentence is processed using algorithm by the data processor. The parser in the system splits the sentence according to pats of speech (POS tagging) such as nouns, adjective, verb, adverb etc. The feature opinion identification identifies the opinions expressed by the reviewers and sentiment it conveys. Thereafter, an output feature opinion pair is provided. This pair categorizes and tells us about the opinion of the reviewer.

Fig 2.2.Review of camera[11]

Customer reviews generally contain the product opinions of many customers expressed in various forms including natural language sentences. A common phenomenon in natural sentence-based customer reviews is that people generally do not express their opinions in a simple way such as “this camera is good,” rather express them using features of the product such as “the battery life of this camera is too short.” Our absolute goal is to search for opinions about features of a target product from a collection of customer review data, analyze the opinion sentences and to determine the orientations of the opinions, and to provide a summary to the user. The Fig 2.2[11] shows user opinion for digital camera. The nature of the review of the customer shows that same sentence can share both the positive and negative aspects about the feature of the product.

The system on feature based opinion mining has applied various methods for feature extraction and refinement including NLP and statistical methods. The system first selects the feature from sentence by considering only information about the term. Secondly, the words such as “picture”, ”image” and “photo” that have similar meaning are considered as different words. Also, some customers give reviews in informal language such “The camera has a gud pic quality” instead of “the camera has a good picture quality”.

In this paper the system will overcome the drawback of informal language reviews segregation into positive and negative. Also, the similar meaning words will be treated as similar feature and grammatical analysis for features will be carried out. Thus, appropriate opinion analysis will take place and appropriate summary will be generated.

III. PREVIOUS WORK

Erik Cambria, Yunqing Xia, Bjorn Schuller , Catherine Havasi(2013) proposed **New Avenues in Opinion Mining and Sentiment Analysis.**

The various opinion mining and sentiment analysis techniques based on NLP and data mining. It also proposed common sentiment analysis task and evolution of opinion mining. Evolution of opinion mining is studied in 3 ways:

- from heuristic to discourse structure
- from coarse to fine-grained analysis
- from keyword to concepts[1]

Heng-Liyang , Qing-Fenglin (2013) proposed **Sentiment Analysis In Multi-scenarios: Using Evolution Strategies For Optimization**

The multi-scenarios problems are .They collected sentences from movie review site and conducted experiment to infer authors’ sentiment .The techniques used for inferring the sentiments of author includes:

- Judgment data collection and preprocessing
- Using ConceptNet to compute characteristic value of sentence
- Optimization weight for multiple scenario [2]

Balla-Müller Nóra, Camelia Lemnaru, Rodica Potolea(2010) proposed **Semi-Supervised Learning with Lexical Knowledge for Opinion Mining**

An implementation of a system that integrates a semi-supervised learning algorithm for text polarity classification

is studied**.** It is a graph based semi-supervised sentiment polarity classifier, where knowledge base is provided for opinion mining. First text classification is performed based on dictionary and then learning. Thereafter, cross-validation is carried out[3].

Vikas Sindhwani and Prem Melville(2008) proposed **Document-Word Co-Regularization for Semi-supervised Sentiment Analysis**

It is a novel semi-supervised sentiment prediction algorithm that utilizes lexical prior knowledge in conjunction with unlabeled examples. It proposes joint sentiment analysis of documents and words based on bipartite graph. Firstly, a prior knowledge of sentiment-laden terms is incorporated. Secondly, we exploit large amount of unlabeled data[5].

Giuseppe Di Fabbrizio, Ahmet Aker, Robert Gaizauska(2011) proposed STARLET: Multi-document Summarization of Service and Product Reviews with Balanced Rating Distributions

A novel approach that uses the rating distribution called as STARLET is used as summarization feature for consistently preserving the overall opinion distribution in the original reviews[6].

Shoushan LI, Chengqing Zong(2008) proposed **Multi-domain Adaptation for Sentiment Classification: using Multiple Classifier Combining Methods**

A multiple classifier system (MCS) is a framework to describe and understand the approach of multi-domain adaptation. Under this framework, a new combining method, called Multi-label Consensus Training (MCT) is proposed. It is used to combine the base classifiers for selecting automatically-labeled samples from unlabeled data in the target domain. This good performance confirms that when data from multiple source domains is available, its combination is an effective way to improve adaptation performance[4].

IV. PROPOSED METHODOLOGY

The proposed system takes care of the informal reviews. The use of English language can be grammatically incorrect with chances of spelling mistakes and wide use of shortcuts. Technically we can phrase these by calling them formal opinions and informal opinions. Where formal opinions refer to the use of proper English with no grammatical mistakes and informal refers to the improper use of English with grammatical mistakes, refers to use of words which are not standard spellings e.g. “4get” instead of “forget”. They say that the UK English is said to be formal English and US English is considered to be informal English. Most of the customers use US English, hence it is very essential to concentrate on these opinions too, or else they might be discarded as noise. Hence the proposed system deals with this aspect of informal reviews.

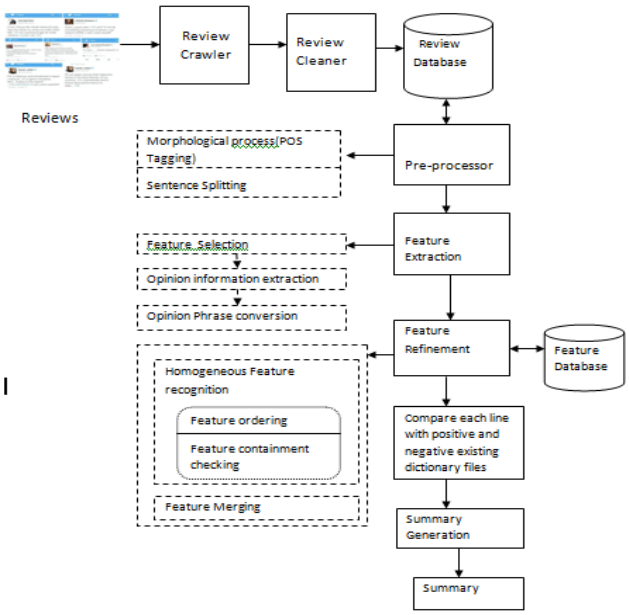

Fig 4.1 FEROM System

To resolve these problems, we propose an enhanced method called, feature extraction and refinement for opinion mining (FEROM). The overall process of FEROM consists of three phases:

- pre-processing,
- feature extraction, and
- feature refinement.

Review Documents Crawler and Review Document Cleaner**:** For a target review site, the crawler retrieves review documents or review data from online stores and stores them locally after filtering mark up language tags. The review cleaner removes unnecessary content such as HTML tags and then stores the review data to the review database.

Document Pre-processor**:** The pre-processor conducts morphological analysis of the review data including POS

tagging, splits a compound sentence into multiple sentences, and performs stop word removal and stemming.

In morphological Analysis the filtered review documents are divided into manageable record-size chunks whose boundaries are decided heuristically based on the presence of special characters. The sentence is scanned according to Parts Of Speech and divided according to that. The compound sentence that are scanned they are divided into multiple simple sentences in sentence splitting. This makes the morphological analysis easy.

Feature Extraction**:** The feature extractor extracts product features from pre-processed review data. Feature extraction proceeds in three phases: In feature selection a candidate feature is selected in a sentence by looking for a noun phrase .After this opinion information extraction finds an opinion phrase that is associated with the candidate feature, and opinion phrase conversion replaces an opinion phrase expressed using a negative term with its antonym.

Feature Refinement**:** This reduces the number of features by merging candidate features with the same or similar meanings, defined as homogeneous features The feature refiner is divided into two parts

In homogeneous feature recognition ,the feature refiner recognizes homogenous features by exploiting the feature ordering process that synchronizes the word orders of the features to detect synonymous feature candidates .The feature containment checking process that examines the subset superset relationship between the features to check for similarity between them.

Feature Merging: Finally, the feature merging process merges homogeneous features into a representative feature and also prunes the feature candidates that have significantly low frequencies and very small amounts of related opinion information.

Comparison: The words and sentences from the feature merging module will be compared with the dictionary of positive and negative words and then the output will be given.

V. EXPERIMENTAL RESULTS

The review documents used in our experiment were taken from a restaurant review site. The training data set and testing data set of the restaurants were taken. The restaurants were rated for five different aspects. The summarization process starts with calculating a set of features for every sentence in training set and testing set. The assumption is that the resulting summary will cover the target sentiments expressed in term of star-ratings by selecting the content that closer mimics the desired distribution and, at the same time, remains within the maximum summary size (100 words).The FEROM method performs the evaluation and it is compared with the othe methods STARLET, MEAD and Random method.The reviews were checked for grammatical errors, redundancy, clarity, coverage and coherence. The chart shows that STARLET outperforms the other two methods and it gives result close to the human reviews that were taken from the manual testing set.The FEROM system gives better result in grammatical check and redundancy checking. The reviews expressed in informal English will be cleaned by FEROM and then the reviews will be segregated into positive and negative.

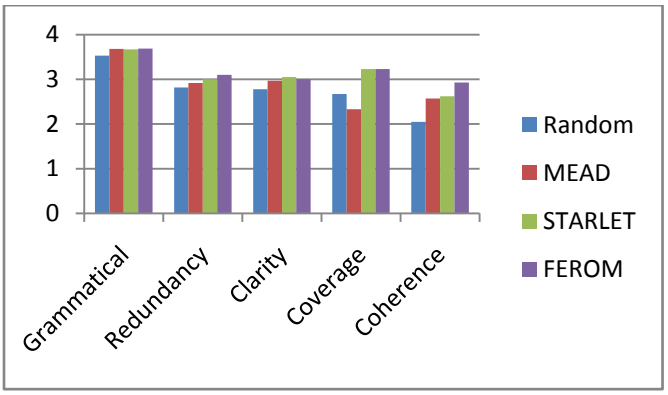

Fig 5.1 A comparison of five aspects on restaurant review.

The redundancy score in FEROM is better than other methods. The informal reviews will be cleaned by FEROM and the output summary will be generated by comparing the output with the inbuilt dictionary.

VI. CONCLUSION

Opinion mining generally refers to the process of extracting product features and opinions from review documents and summarizing them using a graphical representation. Generally, document level opinion mining systems fail to reveal the product features liked or disliked by the users, rather they classify the reviews as positive or negative. A positive review does not mean that the opinion holder has positive opinion on all aspects or features of the product. Similarly, a negative review does not mean that the opinion holder dislikes everything about the product. Keeping in mind the above facts, feature-based opinion mining is

proposed. The proposed method is of feature extraction and refinement for opinion mining, to analyze product review data. It extracts candidate features considering the syntactic and semantic similarities between them and reduces the number of features by merging words with similar meanings.

VII. FUTURE SCOPES

Here the system works on segregation for reviews into positive and negative. Also, it will clean the reviews written in informal English and then the segregation of reviews will be performed. The clarity factor will be improved in future for FEROM. The future work can be done is multilingual comments or reviews that people give on social networking sites.

VIII. ACKNOWLEDGEMENT

I am Dhanashree Raut and I am greatly thankful to Prof. Seema Ladhe for her guidance. We also thank the college authorities, PG coordinator and Principal Sir for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to M.E. Staff and friends.

REFERENCES

[1] Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havas , “New Avenues in Opinion Mining and Sentiment Analysis,”,in IEEE,2013.

[2] Heng-Liyang , Qing-Fenglin , “Sentiment Analysis In Multi-scenarios: Using Evolution Strategies For Optimization,” in Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 2013.

[3] Balla-Müller Nóra, Camelia Lemnaru, Rodica Potolea,“Semi-Supervised Learning with Lexical Knowledge for Opinion Mining ,”in IEEE,2010.

[4] Shoushan LI,Chengqing Zong, “Multi-domain Adaptation for Sentiment Classification: using Multiple Classifier Combining Methods,” in IEEE,2008

[5] Vikas Sindhwani and Prem Melville,”Document-Word Co-Regularization for Semi-supervised Sentiment Analysis,” in IEEE ,2008

[6] G. Di Fabbrizio, A. Aker, and R. Gaizauskas, “STARLET: Multi-Document Summarization of Service and Product Reviews with Balanced Rating Distributions,” in Proc. Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE, 2011, pp. 67–74.

[7] E. Cambria et al., “Semantic Multi- Dimensional Scaling for Open-Domain Sentiment Analysis,” in IEEE Intelligent Systems, preprint, 2012

[8] B. Lu et al., “Multi-Aspect Sentiment Analysis with Topic Models,” in Proc. Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE CS, 2011, pp. 81–88.

[9] D. Olsher, “Full Spectrum Opinion Mining: Integrating Domain, Syntactic and Lexical Knowledge,” in Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, IEEE, 2012, pp. 693–700.

[10] Anton Barhan, Andrey Shakhomirov, “Methods for Sentiment Analysis of Twitter Messages,” in PROCEEDING OF THE 12TH CONFERENCE OF FRUCT ASSOCIATION,2012.

[11] Felipe Jordão Almeida Prado Mattosinho, “Mining Product Opinions and Reviews on the Web,” in Technische Universitat Dresden, 2010.